Comparison of Spline with Kriging in an Epidemiological Problem

Roshanak Alimohammadi

Department of Mathematics,

Alzahra University,

Tehran, Iran.

Abstract

There are various methods to analyze different kinds of data sets. Spatial data is defined when data is dependent on each other based on their respective locations. Spline and Kriging are two methods for interpolating and predicting spatial data. Under certain conditions, these methods are equivalent, but in practice they show different behaviors.

Amount of data can be observed only at some positions that are chosen as positions of sample points, therefore, prediction of data values in other positions is important. In this paper, the link between Spline and Kriging methods is described, then for an epidemiological two dimensional real data set, data is observed in geological longitude and in latitude dimensions, and behavior of these methods are investigated. Comparison of these performances show that for this data set, Kriging method has a better performance than Spline method.

Key Words: *Spatial data, Spline, Kriging.*

1 Introduction

To analyze every kind of data, a model of data structure can be considered. In spatial data analysis, a random field $\{Z(t), t \in D \subset R^d\}$ is applied for spatial data modeling, where t is the site of desired location and D is an index set. For each t, the random field $Z(t)$ can be decomposed as

$$Z(t) = \mu(t) + \delta(t) \tag{1}$$

where $\mu(t)$ is the trend of random field and $\delta(t)$ is a zero mean random field. In spatial statistics, there are many methods for predicting the value of random field at a given spatial site, say t_0, using observations $\mathbf{Z} = (Z(t_1), \dots, Z(t_n))'$ of the random field $Z(.)$ at n spatial sites $t = (t_1, \dots, t_n)$. One of the methods, named Kriging, is the best unbiased predictor which has different kinds such as Ordinary and Universal Krigings. In ordinary kriging, the trend term in relation (1) is fixed and in universal kriging $\mu(t)$ is a function of t (cressie (1993)).

1

Spline is another method for spatial data prediction which minimizes penalized sum of squares criterion. For more details about Splines can refer to Green and Silverman (1994), Hart (2005) and Hardle (2006).

Some authors studied the link between Spline and Kriging, as two prediction methods. Theoretical link between these methods is studied by Kent and Mardia (1994) and the applied link for some data sets is investigated by Hutchinson and Gessler (1994) and Lasslet (1994). In this paper these methods are applied for predicting values of data with two dimensional positions. This data set relates to taberculosis infection prevalence in some cities of Iran which observed in geological longitude and in latitude dimensions. A brief review of Kriging and Spline methods are given respectively in sections 2 and 3, and in section 4 these methods are applied for predicting rate of taberculosis infection prevalence, and performances of the methods are compared. Finally results and conclusions are given in section 5 and the better method to predict values of taberculosis infection prevalence is determined based on the data.

2   Kriging

In Universal Kriging, the trend term in relation (1) is an unknown linear combination of known functions $f_j(.)$ with unknown coefficients β_j, that is

$$\mu(t) = \Sigma_{j=1}^{P+1} \beta_{j-1} f_{j-1}(t)$$

where $\beta = (\beta_0, ..., \beta_p)' \in R^{p+1}$, is an unknown vector of parameters. Furthermore, data $\mathbf{Z}$ can be written as

$$\mathbf{Z} = X\beta + \delta$$

where X is an $n \times (p+1)$ matrix whose (i,j)th element is $f_{j-1}(t_i)$.

It is desired to predict $Z(t_0)$ linearly from data $\mathbf{Z}$, that is

$$\hat{Z}(t_0) = \lambda'\mathbf{Z}, \qquad \lambda'X = x' \tag{2}$$

which is uniformly unbiased $(E[\hat{Z}(t_0)] = E[Z(t_0)])$, and minimizes the mean squares error term $\sigma_e^2 = E[(\hat{Z}(t_0) - Z(t_0))^2]$ over $\lambda = (\lambda_1, ..., \lambda_n)$.

Assumption $\lambda'X = x'$ in equation (2) is equivalent to uniformly unbiased condition, where $x = (f_0(t_0), ..., f_n(t_0))'$. Then the optimal value of λ in relation (2) is

$$\lambda' = [C + X(X'\Sigma^{-1}X)^{-1}(x - X'\Sigma^{-1}C)]' \, \Sigma^{-1} \tag{3}$$

where $C = (c(t_0 - t_1), ..., c(t_0 - t_n))'$ and Σ is an $n \times n$ matrix with (i, j)th element $c(t_i - t_j)$. The Kriging variance can be written as

$$\sigma^2(t_0) = c(0) - C'\Sigma^{-1}C + (x - X'\Sigma^{-1}C)'(X'\Sigma^{-1}X)^{-1}(x - X'\Sigma^{-1}C) \tag{4}$$

When $p = 0$ and $f_0(t) = 1$, universal kriging reduce to ordinary kriging.

In universal kriging, the optimal value of λ (equation (3)) can be written as $\lambda_U = \Sigma_U^{-1} C_U$ where $\lambda_U = (\lambda_1, ..., \lambda_n, -m_0, ..., -m_p)'$ and $m_i s$ are lagrange multipliers that insure $\lambda' X = x'$ and $C_U = (c(t_0 - t_1), ...c(t_0 - t_n), 1, f_1(t_0), ..., f_p(t_0))'$. Then kriging predictor at t_0 is

$$\hat{Z}(t_0) = \mathbf{Z}'_U \Sigma_U^{-1} C_U = V'_U C_U \tag{5}$$

where $V_U = \Sigma_U^{-1} \mathbf{Z}_U$, $\mathbf{Z}_U = (Z(t_1), ..., Z(t_n), 0, ..., 0)'$ which is an $(n+p+1) \times 1$ vector. In equation (5) by writing $V'_U = (V'_1, V'_2)$ so that V_1 is $n \times 1$ and V_2 is $(p+1) \times 1$, then $V_U = \Sigma_U^{-1} Z_U = \begin{bmatrix} \Sigma & X \\ X' & O \end{bmatrix}^{-1} \begin{bmatrix} Z \\ 0 \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$ or $\begin{bmatrix} \Sigma & X \\ X' & o \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} Z \\ 0 \end{bmatrix}$ and dual kriging equations is obtained as

$$\begin{cases} \Sigma V_1 + X V_2 = Z \\ X' V_1 = 0 \end{cases} \tag{6}$$

By solving this system and replacing in relation (2), predictor of $Z(t_0)$ can be written as

$$\hat{Z}(t_0) = V'_1 C + V'_2 \mathbf{x}$$

3   Spline

Data $\mathbf{Z}$ of random field Z(.) is given at locations $\{t_i \in D \subset R^d, d > 1\}$. Consider the problem of estimating unknown function g in the model

$$Z_i = g(t_i) + e_i, \quad i = 1, ..., n \tag{7}$$

To fit g properly, penalized sum of squares criterion is defined as

$$S(g, \lambda) = \Sigma_{i=1}^n (Z_i - g(t_i))^2 + \alpha J_{r+1}^d(g) \tag{8}$$

where $\alpha > 0$ is smoothing parameter. A function $\hat{g}$ which minimizes penalized sum of squares criterion is called Spline. The second term in equation (8) is

$$\begin{aligned} J_{r+1}^d(g) &= \int |\nabla^{r+1} g(t)|^2 dt \\ &= \Sigma_{|m|=r+1} \binom{r+1}{m} \int (\frac{\partial^{r+1} g(t)}{\partial t[1]^{m[1]}, ..., \partial t[d]^{m[d]}})^2 dt \end{aligned}$$

where ∇^{r+1} is $(r+1)$-fold iterated gradient of g, $t = (t[1], ..., t[d])$,

$$J_{r+1}^d(g) = \Sigma_{|m|=r+1} \binom{r+1}{m} \int (\frac{\partial^{r+1} g(t)}{\partial t[1]^{m[1]}, ..., \partial t[d]^{m[d]}})^2 dt$$

where $m = (m[1], ..., m[d])$ and $|m| = m[1] + \cdots + m[d]$.

For $d = 2$, a function $\hat{g}$ which minimizes penalized sum of squares (8) is called Thin Plate spline. To determine a proper value of α can refer to Gu (2002), Hart (2005) and Hardle (2006).

Now finding dual equations for spline in case d=2 is considered (because dimension of our data is d=2.) Smoothing Spline of degree 2 is

$$\hat{Z}(t_0) = a_0 + a_1 x_0 + a_2 y_0 + \Sigma_{i=1}^n b_i e(t_0 - t_i) \tag{9}$$

where

$$e(\mathbf{h}) = ||\mathbf{h}||^2 \log(||\mathbf{h}||^2)/16\pi$$

In relation (9), $a = (a_0, a_1, a_2)'$ and $b = (b_1, \ldots, b_n)'$ solve

$$\begin{cases} K_\alpha \mathbf{b} + X\mathbf{a} = Z \\ X'\mathbf{b} = 0 \end{cases} \tag{10}$$

where $K_\alpha = K + n\alpha I$ is an $n \times n$ matrix with (i,j)th elementary $e(t_i - t_j)$, X is an $n \times 3$ matrix with ith row $(1, x_i, y_i)$, $t_i = (x_i, y_i)'$ and $0 \leq \alpha \leq \infty$ is the smoothing parameter.

4 Application of Spline and Kriging to Prediction

Dual equations (6) and (10) show that the form of these equations for universal kriging and spline are the same, just generalized covariance in Spline is used instead of covariogram. In kriging method, when the second order stationary condition does not satisfied or anyway the IRFk's is used, generalized covariances are applied. Therefore dual equations of kriging and spline methods are equal. Consequently methods of kriging and spline are similar (theoretically), but they can be different practically. In the next section these two methods are compared in an epidemiological problem.

4.1 Data Set and Practical Comparison

Here data of taberculosis infection prevalence in the cities of Iran on the year 1999 is considered. The random field is nonstationary and data has a trend, therefore data is detrended by median polishing. To estimate covariogram, Classic estimator is applied and Gaussian model is chosen as the best model of covariogram for this data set.

To compare the methods performances, a criterion should be considered. Cross validation is a popular means of assessing statistical estimation and prediction. If the variogram model described adequately spatial dependencies implicit in data set, then predicted value $\hat{Z}(t_0)$ should be close to the true value $Z(t_0)$. Ideally additional observations on $Z(.)$ to check this, or initially some of the data might set aside to validate spatial predictor. More likely, all of the data are used to fit the variogram, build the spatial predictor, and there is no possibility of taking more observations. In this case the cross validation approach can be used. Let $2\gamma(h, \hat{\theta})$ be the fitted variogram model (obtained from the data); now delete a datum $Z(t_j)$ and predict it with $\hat{Z}_{-j}(t_j)$ [based on $2\gamma(h, \hat{\theta})$ variogram estimator and the data $\mathbf{Z}$ without $Z(t_j)$].

Its associated mean - square prediction error is $\sigma^2_{-j}(t_j)$ which depends on the fitted variogram model.

The closeness of prediction values to the true values can be characterized as the standardized Mean Square error of Prediction

$$MSP = [1/n(\Sigma_{i=1}^{n} \frac{Z(t_j) - \hat{Z}_{-j}(t_j)}{\sigma_{-j}(t_j)})^2]^{1/2}.$$

In this paper, spline and kriging methods is compared by this criterion and the better method which has smaller MSP is determined. For this data set, gaussian model with nugget effect equal to 39.8 is the best covariogram model to kriging prediction. In spline method the smoothing parameter should be determined and for this data set, the best value which minimizes penalized sum of squares criterion equals $\alpha = 208.6601$.

Cross validation criterion is applied to compare the methods. Programs for computations is written in R and SPLUS environments for the two dimensional data set. Cross validation criterion in kriging method is equal to 0.0239 and in spline method, it is equal to 0.0461.

Consequently, kriging method has a better performance than spline for this data set. This result can be reasonable because in spline usually a special generalized covariance function is used but in kriging this function is characterized based on the data. Therefore for some data sets, kriging method could have better performance than spline.

5   Conclusion

Under certain conditions kriging and spline methods are equivalent, but in practice there are differences between these methods. For instance in spline usually a particular generalized covariance function is used but in kriging, this function is determined based on data, therefore it is expected that kriging has a better performance in some situations. In this paper these methods are applied to predict rate of taberculosis infection prevalence which is a noticeable problem in medicine. The data has measured at two dimensional sites and computations are carried out in R and SPLUS environments. For the data set, computations show that kriging method has a better performance than spline. Consequently application of Kriging can be a preferable method of prediction.

References

[1] Cressie, N. (1993), "Statistics for Spatial Data", New York, John Weily.

[2] Green, P. J. and Silverman, B. W. (1994). "Nonparametric Regression and Generalized Linear Models.",London , Chapman and Hall.

[3] Gu, Ch. (2002), "Smoothing Spline Anova Models", Springer, UK.

[4] Hardle, W. (2006), "Smoothing Techniques: With Implementation in S (Springer Series in Statistics)", Springer, UK.

[5] Hart, J. D. (2005), "Nonparametric Smoothing and Lack-Of-Fit Tests (Springer Series in Statistics)", Springer, UK.

[6] Hutchinson, M. f. and Gessler, F. R. (1994), "Splines - More Than Just a Smooth Interpolator, Geoderma, 62, 45-67.

[7] Kent, J. T. and Mardia, K. V. (1994). "The Link Between Kriging and Thin-Plate Splines", University of Leeds, U.K. , Department of Statistics.

[8] Laslett, G. M. (1994), "Kriging and Splines: An Emprical Comparision of Their Performance in Some Applications", American Statistical Association., Vol. 89, No. 426, 391-400.